Supplementary information S1 (box)

The data sets and analysis used to construct Table 1 are described here. For expression microarrays features were genes. For methylation beadchips, features were probed CpGs. For the SNP chips, used to measure copy number, features were the non-polymorphic probes. For second-generation sequencing, we defined a feature as the coverage (number of mapped reads) in 10,000 basepair windows. For all studies we obtained raw feature level data and performed quantile normaliza- tion [1]. For the SNP chips and gene expression arrays we log-transformed the quantile normalized data. Scripts and data to reproduce our analyses are available from http://rafalab.jhsph.edu/batch

Data set 1

This published data set [2] was demonstrated to have a confounded batch effect [3]. The outcomes of interest were different types of bladder cancer.

Data set 2

This published data set [4] was demonstrated to have a large confounded batch effect [5]. We obtained CEL files, processed with RMA [6]. The outcome of interest was human populations.

Data set 3

This published data set [7] [8] was demonstrated to have a completely confounded batch effect [9]. The outcome of interest was ovarian cancer.

Data set 4

HapMap phase 3 [10] (http://hapmap.ncbi.nlm.nih.gov). The outcome of interest was human populations. We considered only non-related individuals.

Data set 5

This bipolar disease genome wide association study (GWAS) [11] data was obtained from dbGap: http://www.ncbi.nlm.nih.gov/ gap. The dbGaP accession number is phs000017.v3.p1. The data was obtained through the Genetic Association Information Network (GAIN). The outcome of interest was bipolar disease (cases and controls). Note that the publication was related to genotypes and here we analyze copy number data. We do not present evidence that genotype data were affected by batch effects.

Data set 6

These data are from ovarian cancer samples hybridized to Affymetrix HT_HGU133A. All data freely available here: ftp://ftp1. nci.nih.gov/tcga/tumor/ov/cgcc/broad.mit.edu/ht_hg-u133a/transcriptome/. We processed all CEL files with RMA [6].

Data set 7

These data are from ovarian cancer hybridized to Agilent human array. All data freely available here: ftp://ftp1.nci.nih.gov/ tcga/tumor/ov/cgcc/unc.edu/agilentg4502a_07_3/transcriptome/ We used the processed data made available at the url.

Data set 8

Ovarian cancer samples hybridized to Illumina Methylation BeadChip platform. Data freely available here: ftp://ftp1.nci.nih.gov/tcga/tumor/ov/cgcc/jhu-usc.edu/humanmethylation27/methylation/ We used the beta values.

Data set 9

We downloaded aligned reads in chromosome 16 from the 1000 genomes project pilot III data (targeted sequencing of Hapmap samples in 1000-2000 regions, NCBI SRA accession number SRP000033) available from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/. We considered the number of fragments sequenced in 10kb genomic regions as a measure of copy number. We selected samples sequenced with Illumina technology at the Broad Institute and aligned using MAQ [12]. To avoid batch effects due to family, we used a single sample from each identified family in these Hapmap samples. This resulted in looking at 131 different individuals in 6 Hapmap populations. The outcome of interest was human population. We only kept multiple runs for individuals if done on the same date. We binned chromosome 16 into 10Kb regions and use the total number of reads aligned to each bin for each individual as a statistic. For this analysis we kept only regions that overlap exons annotated in Ensembl. These counts were then quantile normalized to get a final count of aligned reads per individual and genomic region.

References

1. Bolstad, B.M., Irizarry, R.A., astrand, M. & speed, t.P. a comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185-93 (2003).
2. Dyrskjot, L. et al. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disre- garding histopathological classification. Cancer Res 64, 4040-8 (2004).
3. Zilliox, M.J. & Irizarry, R.A. a gene expression bar code for microarray data. Nat Methods 4, 911-3 (2007).
4. Spielman, R.S. et al. common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39,
226-31 (2007).

5. Akey, J.M., Biswas, S., Leek, J.T. & Storey, J.D. on the design and analysis of gene expression studies in human populations. Nat
Genet 39, 807-8; author reply 808-9 (2007).
6. Irizarry, R.A. et al. exploration, normalization, and summaries of high density oligonucleotide array probe level data.
Biostatistics 4, 249-64 (2003).
7. Petricoin, E.F. et al. use of proteomic patterns in serum to identify ovarian cancer.
Lancet 359, 572-7 (2002).
8. Conrads, T.P. et al. High-resolution serum proteomic features for ovarian cancer detection. Endocr Relat Cancer 11, 163-78
(2004).
9. Baggerly, K.A., Edmonson, S.R., Morris, J.S. & Coombes, K.R. High-resolution serum proteomic patterns for ovarian cancer
detection. Endocr Relat Cancer 11, 583-4; author reply 585-7 (2004).
10. Dick, D.M. et al. Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the national Institute of
Mental Health Genetics Initiative. Am J Hum Genet 73, 107-14 (2003).
11. Hapmap. the International HapMap Project. Nature 426, 789-96 (2003).
12. Li, H., Ruan, J. & Durbin, R. Mapping short Dna sequencing reads and calling variants using mapping quality scores. Genome
Res 18, 1851-8 (2008).